Model for count data

Although Poisson model is useful, sometimes the assumptions of the Poisson model are violated.

Two extensions: Zero inflated Poisson and Negative Binomial regression model

Zero inflated Poisson regression model:

What is zero inflated and why?

Model:

Example:

#Source: http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm

#quoted from http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm

https://stats.idre.ucla.edu/stata/dae/zero-inflated-poisson-regression/

**"Example 2**. The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors are asked how long they stayed, how many people were in the group, were there children in the group and how many fish were caught. Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data.

require(ggplot2)

require(pscl)

require(boot)

#zinb <- read.csv("http://www.ats.ucla.edu/stat/data/fish.csv")

**zinb<-read.csv("J://fish.csv")**

zinb <- within(zinb, {

nofish <- factor(nofish)

livebait <- factor(livebait)

camper <- factor(camper)

})

summary(zinb)

ggplot(zinb, aes(count)) + geom\_histogram() + scale\_x\_log10()

安裝以及載入pscl package

比較以下model

zero-inflated Poisson model:

summary(m11 <- zeroinfl(count ~ child + camper | persons, data = zinb))

summary(m1 <- zeroinfl(count ~ child + camper+persons | child+camper+persons, data = zinb))

Original Poisson model:

summary(p1 <- glm(count ~ child + camper+persons, family = poisson, data = zinb))

overall goodness of fit test for Poisson model

with(p1, cbind(res.deviance = deviance, df = df.residual,

p = pchisq(deviance, df.residual, lower.tail=FALSE)))

Vuong non-nested hypothesis testing to compare different models:

vuong(p1, m1)

#vuong(mll, m1)

Model:

Likelihood

How to explain the results?

PSCL package manual:

https://cran.r-project.org/web/packages/pscl/pscl.pdf

Over dispersion:

One of the important assumptions of the Poisson model is equi-dispersion. That is,

the mean and variance are equal:

One way to solve the over dispersion problem is to use an alternative distribution for count data.

Negative Binomial regression model:

Negative Binomial Model

The Negative Binomial Distribution

independent Bernoulli trials with success rate p

X~NB(r , p )

X:# of successes before r failures (r>0)

P(X=k)=C(k+r-1,k)(1-p)^r\*p^k k=0,1,2,...

when r is not a positive integer use Gamma function to express P(X=k)

ƛ=E(X)= p r /(1-p)

Var(X)= p r /(1-p)^2

Var(X)=cE(X), c=?

the relation between p and ƛ ƛ =p r/(1-p), p= ƛ/( ƛ+ r)

For regression model, let data be (Yi,Xi)

E(Yi|Xi)= ƛi=exp(Xi'b), where b is the regression coefficient vector.

(pi= ƛi/( ƛi+ r))

Var(Yi|Xi)=

Var(Yi|Xi)/E(Yi|Xi)=1/(1-pi)=1+ (ƛi/r)

So Var(Yi|Xi)= E(Yi|Xi)( 1+ (ƛi/r))= E(Yi|Xi)( 1+ (E(Yi|Xi)/r)). 1/r is the overdispersion parameter

Likelihood is a function of p and b

Dispersion parameter

R code for Negative Binomial regression:

Source: http://www.ats.ucla.edu/stat/r/dae/nbreg.htm

require(foreign)

require(ggplot2)

require(MASS)

#dat <- read.dta("http://www.ats.ucla.edu/stat/stata/dae/nb\_data.dta")

library(foreign)

dat <- **read.dta**("https://stats.idre.ucla.edu/stat/stata/dae/nb\_data.dta")

#new source: https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/

Example 1. School administrators study the attendance behavior of high school juniors at two schools. Predictors of the number of days of absence include the type of program in which the student is enrolled and a standardized test in math.

We have attendance data on 314 high school juniors from two urban high schools in the file nb\_data. The response variable of interest is days absent, daysabs. The variable math gives the standardized math score for each student. The variable prog is a three-level nominal variable indicating the type of instructional program in which the student is enrolled.

## save file nb\_data.dta in統計諮詢\_報告

#dat<-read.dta("J:/nb\_data.dta")

#dat<-read.dta("I:/統計諮詢\_報告/negativebinomial\_data.dta")

dat<-read.dta(“D:/統計諮詢2020fall/牙周病分析/negativebinomial\_data.dta”)

dat <- within(dat, {

prog <- factor(prog, levels = 1:3, labels = c("General", "Academic", "Vocational"))

id <- factor(id)

})

summary(dat)

ggplot(dat, aes(daysabs, fill = prog)) +

geom\_histogram(binwidth=1) +

facet\_grid(prog ~ ., margins=TRUE, scales="free")

with(dat, tapply(daysabs, prog, function(x) {

sprintf("M (SD) = %1.2f (%1.2f)", mean(x), sd(x))

}))

summary(m1 <- glm.nb(daysabs ~ math + prog, data = dat))

Checking goodness of fit for Poisson regression model

m3 <- glm(daysabs ~ math + prog, family = "poisson", data = dat)

X2 <- 2 \* (logLik(m1) - logLik(m3))

X2

pchisq(X2, df = 1, lower.tail=FALSE)

newdata1 <- **data.frame**(math = **mean**(dat$math), prog = **factor**(1:3, levels = 1:3, labels = **levels**(dat$prog)))

newdata1$phat <- **predict**(m1, newdata1, type = "response")

newdata1

newdata2 <- data.frame(

math = rep(seq(from = min(dat$math), to = max(dat$math), length.out = 100), 3),

prog = factor(rep(1:3, each = 100), levels = 1:3, labels =

levels(dat$prog)))

newdata2 <- cbind(newdata2, predict(m1, newdata2, type = "link", se.fit=TRUE))

newdata2 <- within(newdata2, {

DaysAbsent <- exp(fit)

LL <- exp(fit - 1.96 \* se.fit)

UL <- exp(fit + 1.96 \* se.fit)

})

ggplot(newdata2, aes(math, DaysAbsent)) +

geom\_ribbon(aes(ymin = LL, ymax = UL, fill = prog), alpha = .25) +

geom\_line(aes(colour = prog), size = 2) +

labs(x = "Math Score", y = "Predicted Days Absent")

Explain the results:

For each one-unit increase in math, the expected count of the number of days absent in log scale decreases by 0.006

Homework:

Response variable: 拔牙顆數 分三期(治療前 中 後) 分別作分析

covariates: 年紀(continuous), 性別(0,1),

以及ses (social economic status, categorical variable) and area (categorical variable)

請用

1. Poisson Regression

2. Zero-inflated Poisson Regression

3. Negative Binomial Regression

4. Zero-inflated Negative Binomial Regression

來作分析

Zero inflated Negative Binomial Regression

Model:

Likelihood

install.packages("ggplot2")

install.packages("pscl")

install.packages("MASS")

install.packages("boot")

install.packages("foreign")

require(ggplot2)

require(pscl)

require(MASS)

require(boot)

require(foreign)

zinb <- read.csv("https://stats.idre.ucla.edu/stat/data/fish.csv")

zinb <- within(zinb, {

nofish <- factor(nofish)

livebait <- factor(livebait)

camper <- factor(camper)

})

summary(zinb)

ggplot(zinb, aes(count, fill = camper)) +

geom\_histogram() +

scale\_x\_log10() +

facet\_grid(camper ~ ., margins=TRUE, scales="free\_y")

m1 <- zeroinfl(count ~ child + camper | persons,

data = zinb, dist = "negbin", EM = TRUE)

summary(m1)

summary(m2 <- glm.nb(count ~ child + camper, data = zinb))

vuong(m1, m2)